http://dx.doi.org/10.1111/j.1467-842X.2006.00445.x

Clarke, B.R. and Schubert, D.D. (2006) An adaptive trimmed likelihood algorithm for identification of multivariate outliers. Australian & New Zealand Journal of Statistics, 48 (3). pp. 353-371.

REVISED PAPER SENT TO ANZJS

An Adaptive Trimmed Likelihood Algorithm for Identification of Multivariate Outliers

BRENTON R. CLARKE[1*] and DANIEL D. SCHUBERT[1]

Murdoch University

Summary

This article describes an algorithm for the identification of outliers in multivariate data based on the asymptotic theory underlying the trimmed likelihood estimator and the minimum covariance determinant estimator (MCD). The strategy is, to choose a subset of the data which minimizes an appropriate measure of the asymptotic variance of the multivariate location estimator. Observations not belonging to this subset are considered outliers which are to be trimmed. We take as the correct trimming proportion, α less than approximately 50%, the minimum of any minima of this asymptotic variance occurring for an $\alpha > 0$. If no local minima occur for an $\alpha > 0$ then the data set will be considered outlier free.

Key words: Trimmed Likelihood Estimator; Minimum Covariance Determinant Estimator; Asymptotic Variance; Mahalanobis Distance; Forward Search Algorithm; Adaptive Estimation.

1 Introduction

The proposal for the identification of multivariate outliers, described in this paper, finds its origins in the discussion of the trimmed likelihood estimator developed in Bednarski

[1*] Author to whom correspondence should be addressed.

[1] Mathematics and Statistics, Division of Science and Engineering, Murdoch University, Murdoch, Western Australia, 6150

[1*] E-mail: B.Clarke@murdoch.edu.au

& Clarke (1993) and in consequent discussion of adaptive versions of the estimator. For example see Clarke (1994, 2000), Clarke et al. (2000) and Bednarski & Clarke (2002). An adaptive trimmed likelihood algorithm (ATLA) has been used to identify outliers in univariate data sets whilst yielding parameter estimates no longer impacted by any outlying data. ATLA has also been used to simultaneously identify outliers in regression and give the least squares estimator with the outliers removed.

Most outliers, in a univariate setting, can be exposed by a simple scatter diagram or a stem-leaf plot but can still upset estimation when one uses non-visual methods of assessment. Clusters of contaminant data can confuse even the visual perspective. Outliers can have a disastrous impact upon Least Squares regression and inference which instigated, for example, the Least Trimmed Squares, LTS (Butler 1982, Rousseeuw 1983, 1984) and Least Median of Squares, LMS (Rousseeuw 1983, 1984 and Rousseeuw & Leroy 1987) approaches to regression.

With multivariate estimation, visual inspections are not possible in greater than 3 dimensions. Even 3 dimensional data sets can consist of outliers that evade detection from certain angles and so one is dependent on analytical methods for outlier detection.

A cluster of outliers distributed about a point displaced from the true mean and with a similar covariance structure to the clean data, shift outliers (Rocke & Woodruffe 1993), can lead to poor estimation and inference. Other perhaps more easily identified contaminants, with similar consequences for statistical inference, are the solitary strays or a scatter of stray points, termed linear and radial outliers.

For an assumed normal distribution the univariate location or regression estimator defined as a trimmed likelihood estimator corresponds to the LTS estimator, but with a trimming proportion thought to be other than approximately 50%. Adaptive trimming, such as in ATLA, seeks to choose a trimming proportion, $0 \leq \alpha < \frac{1}{2}$, so that α minimizes a suitable estimate associated with the asymptotic variance of the trimmed likelihood estimator. In multivariate estimation the corresponding estimator to the LTS is the minimum covariance determinant (MCD) estimator (Rousseeuw 1983, Butler et al 1993). Therefore, assuming a multivariate normal distribution, the equivalent of ATLA is to choose that multivariate MCD estimate for location with a trimming proportion that minimizes the estimate of the determinant of the asymptotic variance of the MCD estimate, trimming possibly less than approximately 50%.

The new proposal outlined in this paper again involves starting with a robust MCD estimate for centroid and scale followed by a Forward Search Algorithm (Hadi 1992, 1994, Aktinson 1994, Rocke & Woodruffe 1996) which searches for a subset of data resulting in the minimum determinant of the estimated asymptotic variance. The theoretical premise for our proposal is based on the principle that trimming a normally distributed data set will necessarily reduce the information it contains. This reduction in information increases the estimated variability of the location estimate unless those observations trimmed are outliers.

This proposal is unique in that the statistical analyst is not seeking an *outlier region* (Becker & Gather 1999) proper but is rather targeting existing observations as necessarily belonging to this region without exactly specifying any orthodox *cut-off* value. Section 2 will describe the proposal while section 3 will address its performance assessing simulated data. Section 4 will discuss comparison with other existing outlier detection methodology and section 5 provides a series of illustrations of the proposal when applied to real data sets.

2 Proposal

Consider any set of independent identically distributed observations $\{\boldsymbol{X}_i\}_{i=1}^{n}$ in p-dimensional Euclidean space $\Re^p$ and let S_n be an arbitrary subset of size $s_n = \lfloor n\gamma \rfloor$ for $\gamma = 1-\alpha$ where α corresponds to the amount of trimming, $0.5 < \gamma \leq 1$. As in the previous literature cited we now denote $T[F_n]$ to be the univariate estimate of location and then correspondingly

$\boldsymbol{T}[F_n]$ to be the multivariate estimate of location. Here $\boldsymbol{T}[\cdot]$ is the estimating functional and F_n is the empirical distribution assigning atomic mass $\frac{1}{n}$ to each of the points $\boldsymbol{X}_i$ in the sample. In the univariate location estimation there is according to Rousseeuw (1983), Butler (1982) and Bednarski & Clarke (1993), an asymptotic normality result, which holds for symmetric distributions,

$$\sqrt{n}(T[F_n] - \mu) \xrightarrow{d} N(0, V(\alpha, F))$$

where the asymptotic variance equals

$$V(\alpha, F) = \frac{\int_{-x_\alpha}^{x_\alpha} x^2 dF_0(x)}{\{1 - \alpha - 2x_\alpha f_0(x_\alpha)\}^2} \ .$$

Here $x_\alpha = F_0^{-1}(1 - \alpha/2)$ where F_0 is the underlying distribution F that is centred at zero and has density f_0.

That subset of univariate observations, corresponding to an $\alpha \geq 0$, which minimizes $V(\alpha, F)$ would be considered free of outliers. Hence to estimate and minimize $V(\alpha, F)$ assuming an underlying normal distribution we minimize

$$\frac{\frac{1}{\lfloor n\gamma \rfloor} \sum_{i \in S_n} (x_i - \bar{x})^2}{\{1 - \alpha - \sqrt{\frac{2}{\pi}} z_{\alpha/2} e^{-\frac{1}{2} z_{\alpha/2}^2}\}^2} \tag{1}$$

where $z_{\alpha/2}$ is the appropriate critical point of the standard normal distribution. This is equivalent to minimizing

$$\nu(\alpha, F_n) = \frac{\bar{\sigma}_\alpha^2[F_n]}{\{1 - \alpha - \sqrt{\frac{2}{\pi}} z_{\alpha/2} e^{-\frac{1}{2} z_{\alpha/2}^2}\}^2} \ .$$

See Bednarski & Clarke (1993, 2002). Potential outliers are those observations in the sample not in the set S_n which minimize the above two expressions. Note here α is chosen from a set $\{0, \frac{1}{n}, \frac{2}{n}, \ldots, \frac{\lfloor n/2 \rfloor}{n}\}$.

When analyzing multivariate data sets we are dealing with elliptical probability densities of the form

$$\frac{1}{|\boldsymbol{\Sigma}|^{1/2}} f((\boldsymbol{x} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1} (\boldsymbol{x} - \boldsymbol{\mu})) \tag{2}$$

where $f : \Re^+ \to \Re^+$ is assumed to be nonincreasing, yielding a unimodal density (Butler et al 1993), with $\boldsymbol{\mu}$ our centroid, $\boldsymbol{\Sigma}$ our covariance matrix and $|\boldsymbol{\Sigma}|$ indicating the determinant of $\boldsymbol{\Sigma}$ which is assumed to be non zero.

If the sample data is from a multivariate normal distribution with mean zero and covariance matrix being the identity, $N(\mathbf{0}, \boldsymbol{I}_p)$, then the sample covariance matrix using the MCD estimator of Butler et al. (1993) is such that

$$\hat{\boldsymbol{\Sigma}}_\alpha[F_n] \xrightarrow{wp1} \rho(\gamma)\boldsymbol{I}_p.$$

Given the normalization required to ensure the integral of (2) is 1, and transforming to polar co-ordinates (Davies 1987, Butler et al 1993), we arrive at

$$\rho(\gamma) = \int_E \boldsymbol{x}\boldsymbol{x}^\top dF(\boldsymbol{x}) = \frac{2\pi^{p/2}}{p\Gamma(p/2)} \int_0^{\sqrt{\chi^2_{1-\alpha,p}}} r^{p+1}\phi(r^2)dr \tag{3}$$

where the first integral is over the set $E = \{\boldsymbol{x} | (\boldsymbol{x} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{x} - \boldsymbol{\mu}) \leq r_\gamma^2\}$ for an r_γ^2 chosen so that $F\{E\} = \gamma$.

If in fact the data are multivariate normal, with known mean and covariance matrix, then it is well known that $r_\gamma^2 = \chi^2_{1-\alpha,p}$. Here $\chi^2_{1-\alpha,p}$ is the critical point of a chi squared distribution with p degrees of freedom corresponding to the dimension of our data and having $1 - \alpha$ area under the chi squared density curve to the left of it. Γ is the usual gamma function, $\Gamma(v) = \int_0^\infty s^{v-1}e^{-s}ds$.

With regard to our sample estimate of the multivariate mean, $\boldsymbol{T}[F_n]$, we have

$$\sqrt{n}(\boldsymbol{T}[F_n] - \boldsymbol{\mu}) \xrightarrow{d} N(\mathbf{0}, \kappa(\gamma)\boldsymbol{I}_p)$$

where (see Butler et al. 1993)

$$\kappa(\gamma) = \frac{p\Gamma(p/2)\int_0^{r_\gamma} r^{p+1}\phi(r^2)dr}{8\pi^{p/2}(\int_0^{r_\gamma} r^{p+1}\phi'(r^2)dr)^2} = \frac{\rho(\gamma)}{(\frac{4\pi^{p/2}}{p\Gamma(p/2)}\int_0^{r_\gamma} r^{p+1}\phi'(r^2)dr)^2} \quad . \tag{4}$$

Here again $r_\gamma = \sqrt{\chi^2_{1-\alpha,p}}$ and $\phi(u) = (1/(2\pi))^{p/2}e^{-u/2}$ whence substituted in for f in (2) leads to the multivariate normal distribution.

If our data was exactly from a multivariate normal the $\kappa(\gamma)$ above can be used to give an estimate for the asymptotic variance of $\boldsymbol{T}[F_n]$ assuming $\gamma = 1 - \alpha$ where α is the proportion of trimming. If our sample data consists of outliers then one would expect the value of $\rho(\gamma)\boldsymbol{I}_p$ to disagree with our sample variance $\hat{\boldsymbol{\Sigma}}_\alpha[F_n]$ which is no longer a covariance matrix from a normal distribution. With this in mind the multivariate extension of minimizing (1) becomes as a direct consequence of (4), choose γ (equivalently choose α) to minimize:

$$|\frac{\kappa(\gamma)}{\rho(\gamma)}\hat{\Sigma}_\alpha[F_n]| = \frac{|\hat{\Sigma}_\alpha[F_n]|}{(\frac{4\pi^{p/2}}{p\Gamma(p/2)}\int_0^{r_\gamma} r^{p+1}\phi'(r^2)dr)^{2p}} \tag{5}$$

In fact the above formula (5), which we will call Type 1, is equivalent to minimizing $V(\alpha, F_n)$ when $p = 1$ which is the preferred option 4 in Clarke (1994).

By Bednarski & Clarke (1993) for univariate data, $p = 1$ and $F = \Phi$, the cumulative standard normal distribution, we see that

$$(1-\alpha)\hat{\Sigma}_\alpha[F_n] \xrightarrow{wp1} \int_{-z_{\alpha/2}}^{z_{\alpha/2}} x^2 d\Phi(x).$$

This yields the Fisher consistent estimate for $\boldsymbol{\Sigma}$ which in the multivariate setting would result in choosing γ to minimize, our Type 2 proposal,

$$|\frac{\kappa(\gamma)}{\rho(\gamma)}(1-\alpha)\hat{\Sigma}_\alpha[F_n]| = \frac{|(1-\alpha)\hat{\Sigma}_\alpha[F_n]|}{(\frac{4\pi^{p/2}}{p\Gamma(p/2)}\int_0^{r_\gamma} r^{p+1}\phi'(r^2)dr)^{2p}}\,. \tag{6}$$

For $p = 1$ minimizing the objective function immediately above, (6), is equivalent to choosing option 5 in Clarke (1994). We explain the performance of the Type 1 and Type 2 proposals later in this paper.

The initial step for our proposal involves obtaining a robust MCD estimate for the centroid of the sample data set S_n using an algorithm based on Rousseeuw & Leroy (1987) and Woodruff & Rocke (1993). This algorithm is needed to calculate the MCD estimate for location since for multivariate data sets of dimension p and size n a combinatorial scale of trials is required to find that subset of size $\lfloor(n+p+1)/2\rfloor$ yielding the minimum covariance determinant. This entails the consideration of $\frac{n!}{k!(n-k)!}$ subsets where $k = (\lfloor(n+p+1)/2\rfloor)$. For samples of size $n = 20$ this equates to 167 960 subsets to be investigated when $n = 50$ this has already increased to $> 1.2 \times 10^{14}$ thus the computer time needed to locate the MCD centre becomes absurdly large.

Once our MCD estimate for location and scale has been evaluated the procedure is to initially arrange each sample point in ascending order of Mahalanobis distance from this

MCD estimate for centroid $\tilde{\boldsymbol{X}}$:

$$M_i = \sqrt{(\boldsymbol{X}_i - \tilde{\boldsymbol{X}})^\top \hat{\Sigma}^{-1} (\boldsymbol{X}_i - \tilde{\boldsymbol{X}})} \tag{7}$$

where $\hat{\Sigma}$ is the covariance of those observations contributing to the MCD estimate. The data arranged thus, $\{M_1, M_2, \ldots, M_h, M_{h+1}, \ldots, M_n\}$ where $h = \lfloor \frac{n+p+1}{2} \rfloor$ for p-dimensional data is then divided into two subsets, first the set of h points closest to the centroid, that is those points responsible for the MCD estimate, and a set of $n - h$ points being tested for outlyingness. Once the objective function being used, either (5) or (6), to assess this first subset, the subset is inflated to include the nearest point, that is that observation in the complement of the subset closest to the subset, whence the whole sample is again arranged in ascending order of Mahalanobis distance using the centroid and covariance matrix derived from this inflated subset. With this step, and each subsequent repetition thereof, it is imperative to note that some members of the *previously* assessed subset can interchange with complement members *before* a further inflation due to the re-ordering. (Atkinson, Riani, Cerioli 2004 pg 306). The objective function being used is again assessed for this new subset before the subset is again incremented to include the nearest observation. This procedure is repeated until the subset has been inflated to include the whole sample set. Ideally that subset yielding the minimum for the objective function being employed will be regarded as outlier-free, so observations, if any, not a member of this subset are identified as outlying.

In the next section we see that it is more appropriate to choose that subset corresponding to the *minimum* of any minima occurring for an $\alpha > 0$ as outlier free. In the event no local minima occurs for an $\alpha > 0$ then the data set is considered outlier free.

3 Monte Carlo results

We conducted Monte Carlo simulations for both 2 and 4 dimensional data sets to assess the efficacy of our Type 1 and Type 2 proposals. Data sets distributed normally, then contaminated with a pre-specified proportion, ϵ, of outliers were subjected to the new proposal with the average proportion of outliers detected and trimming frequency recorded. Sample sizes of $n = 20,\ 50,\ 100,\ 500$ were generated with a proportion $\epsilon = 0,\ 1/n,\ 0.3,\ 0.4$ of the data shifted from the main mean.

For the contaminated data, with $\epsilon = 1/n,\ 0.3$, this outlying proportion of data consisted of a pth variable centred about a displacement d from the *clean* data distributed $N([0,0], \boldsymbol{I_2})$, $N([0,0,0,0], \boldsymbol{I_4})$, where $d = q\sqrt{\chi^2_{0.975,2}},\ q\sqrt{\chi^2_{0.975,4}}$ for $q = 2,\ 4$ (Juan & Prieto 2001),

1. for $q = 2$ we have $N([0, 5.4324]^\top, \boldsymbol{I_2})$ and $N([0,0,0,6.6763]^\top, \boldsymbol{I_4})$
2. for $q = 4$ we have $N([0, 10.8348]^\top, \boldsymbol{I_2})$ and $N([0,0,0,13.3526]^\top, \boldsymbol{I_4})$

When assessing the data sets with a proportion $\epsilon = 0.4$ of contamination, these outliers formed two clusters of outlying data, each cluster representing $0.2n$ of the whole data set. These two clusters were centred about a displacement of $\pm d$ respectively.

Table 1 contains, for both 2 and 4 dimensional data sets, the sample size n, the proportion of planted outliers ϵ and the average displacement, d, of the outlying variable. The proportion of instances, for an $\alpha > 0$, when there was at least one subset of data forcing a minima of our objective function was recorded, p_t, along with the average trimming proportion $\bar{\alpha}$.

When generating clean data sets, $\epsilon = 0$, we discovered that the Type 2 proposal was too sensitive for sample sizes $n \leq 30$ but performed better for larger samples. With this insight we decided to apply the proposal Type 1 to samples $n = 20$ and use Type 2 for the other Monte Carlo sample sizes $n \geq 30$.

The *clean* data sets should yield a solitary minimum at $\alpha = 0$, indicating no outliers. Our Monte Carlo trials show that as the sample size increases, even to as low as $n = 50$, in less than 1/1000 cases did any minima occur for an $\alpha > 0$. The identifying of *good* data as outlying happened extremely rarely for both 2 and 4 dimensional data sets. For $n = 500$ there was never any reported instances of minima occurring for an $\alpha > 0$, thus no data was ever identified as outlying when the data sets of size $n = 500$ were *not* corrupted with outliers, as should be the case.

When a single outlier was planted, $\epsilon = 1/n$, and distributed $N([\mathbf{0}, 4\sqrt{\chi^2_{0.975,p}}], \boldsymbol{I_p})$ for $p = 2,\ 4$ and $\mathbf{0}$ representing the zero vector of length $p-1$, we achieved excellent results with a global minimum occurred for $\alpha = 1/n$ in excess of 99% of occasions. When a single outlier was planted $N([\mathbf{0}, 2\sqrt{\chi^2_{0.975,p}}], \boldsymbol{I_p})$ it was less likely to be identified for small samples which can be expected when one considers the effect of *chaining* (Wilks 1995) and its masking impact when using a distance based algorithm.

For a single outlying cluster, $\epsilon = 0.3$, we again see that for the larger average displacement, $d = 4\sqrt{\chi^2_{0.975,p}}$, in over 99% of occasions very close to the correct proportion of outliers was trimmed if not the exact proportion. This outlying cluster was identified by a local minimum occurring for $\alpha > 0$, in this case $\alpha \approx 0.3$.

For 2 dimensional data sets of size $n \geq 50$, when the average displacement of the contaminated variable for the cluster was $d = 2\sqrt{\chi^2_{0.975,p}}$, the cluster was identified on at least 90% of occasions by the Type 2 proposal. In the case of 4 dimensional data the figures are even better and for sample sizes $n = 500$ the cluster of outliers is nearly always identified. Table 1 also shows the results for data sets consisting of two outlying clusters, $\epsilon = 0.4$, were equally as encouraging.

When confronted with data sets contaminated by outlying clusters we found increasing instances of multiple minima for $\alpha > 0$. The larger the proportion of clustered outliers, the more frequent the occurrence of ~~multiple~~ minima. Table 2 contains the proportion of such instances, p_M, for an $\epsilon = 0.1$, and0.3 respectively, for samples of size $n = 50,\ 100$ in both 2 and 4 dimensional cases.

Fixed before submission

We can see, for example, that for bivariate data sets, whence $\epsilon = 0.3$ and $d = 2\sqrt{\chi^2_{0.975,2}}$, there occurred multiple minima for $\alpha > 0$ for between 15-20% of data sets. The strategy

Table 1: Simulation results

dimension p=2					*dimension p=4*				
n	ϵ	d	p_t	$\bar{\alpha}$	n	ϵ	d	p_t	$\bar{\alpha}$
20	0		0.072	0.0248	20	0		0.066	0.0228
	0.05	5.4324	0.66	0.0386		0.05	6.6763	0.479	0.0374
		10.8648	> 0.999	0.0513			13.3526	> 0.999	0.0506
	0.3	5.4324	0.641	0.2001		0.3	6.6763	0.542	0.1699
		10.8648	> 0.999	0.3083			13.3526	0.992	0.3027
	0.4	5.4324	0.587	0.2384		0.4	6.6763	0.372	0.1438
		10.8648	> 0.999	0.406			13.3526	0.941	0.3752
50	0		0.011	0.0033	50	0		0.003	0.0009
	0.02	5.4324	0.716	0.0158		0.02	6.6763	0.775	0.0161
		10.8648	> 0.999	0.0207			13.3526	> 0.999	0.0202
	0.3	5.4324	0.912	0.2799		0.3	6.6763	0.925	0.283
		10.8648	> 0.999	0.3104			13.3526	> 0.999	0.3058
	0.4	5.4324	0.909	0.3713		0.4	6.6763	0.904	0.3692
		10.8648	> 0.999	0.4116			13.3526	> 0.999	0.4089
100	0		0.001	< 0.0001	100	0		0.002	0.0003
	0.01	5.4324	0.705	0.0072		0.01	6.6763	0.81	0.0101
		10.8648	> 0.999	0.0101			13.3526	> 0.999	0.0102
	0.3	5.4324	0.958	0.2912		0.3	6.6763	0.987	0.2997
		10.8648	> 0.999	0.3077			13.3526	> 0.999	0.3042
	0.4	5.4324	0.963	0.3889		0.4	6.6763	0.979	0.3961
		10.8648	> 0.999	0.4083			13.3526	> 0.999	0.4056
500	0		0.001	< 0.0001	500	0		< 0.001	< 0.0001
	0.002	5.4324	0.509	0.001		0.002	6.6763	0.732	0.0015
		10.8648	> 0.999	0.002			13.3526	> 0.999	0.002
	0.3	5.4324	> 0.999	0.3032		0.3	6.6763	> 0.999	0.3004
		10.8648	> 0.999	0.3061			13.3526	> 0.999	0.3008
	0.4	5.4324	0.999	0.4039		0.4	6.6763	> 0.999	0.4035
		10.8648	> 0.999	0.4072			13.3526	> 0.999	0.4039

Clean data sets, $\epsilon = 0$. Solitary outlier, $\epsilon = 1/n$. One outlying cluster, $\epsilon = 0.3$.

Two outlying clusters, $\epsilon = 0.4$.

Table 2: Frequency of Multiple Minima

dimension p=2				*dimension p=4*			
n	ϵ	d	p_M	n	ϵ	d	p_M
50	0		0	50	0		0
	0.02	5.4324	0.003		0.02	6.6763	0.004
		10.8648	0.008			13.3526	0.004
	0.1	5.4324	0.091		0.1	6.6763	0.035
		10.8648	0.042			13.3526	0.045
	0.3	5.4324	0.153		0.3	6.6763	0.087
		10.8448	0.101			13.3526	0.073
100	0		0	100	0		0
	0.01	5.4324	< 0.001		0.01	6.6763	< 0.001
		10.8648	< 0.001			13.3526	< 0.001
	0.1	5.4324	0.17		0.1	6.6763	0.069
		10.8648	0.023			13.3526	0.013
	0.3	5.4324	0.168		0.3	6.6763	0.043
		10.8448	0.066			13.3526	0.024

is therefore, to select that particular $\alpha > 0$ which results in the *minimum* of *any* minima satisfying $\alpha > 0$. Notice as the dimension increases the instances of multiple minima for $\alpha > 0$ diminished. With 4 dimensional data sets corrupted by a proportion $\epsilon = 0.3$ of data displaced from the main mean, multiple minima occurred on less than 10% of occasions.

For a solitary outlier, $\epsilon = 1/n$, very few instances of multiple minima were reported for sample sizes $n = 50$ and none for $n = 100$ for both the displacement sizes chosen for d.

The proposal Type 2 was applied to bivariate data sets of size $n = 100$ with proportions $\epsilon = 0.1,\ 0.3$ corrupted with a shifted mean of $d = 4\sqrt{\chi^2_{0.975,2}}$. Collected, for Figures 1-4, were instances of multiple minima occurring for $\alpha > 0$ and the relationship between the size of the retained subset and the corresponding value of our objective function (6) are illustrated.

Figures 1 and 2 depict a steep fall in our objective function (6), in the vicinity of the correct trimming proportion; note these plots represent cases of multiple minima for $\alpha > 0$ *only*. For the majority of cases only one minima occurs for $\alpha > 0$ and the plots (not shown) for these cases depict the same dramatic fall in the value of (6).

Closer inspections, on reduced intervals of subset size for clarity with $d = 4\sqrt{\chi^2_{0.975,2}}$, see Figures 3 and 4, show more examples of multiple minima but ensures us that if we apply the strategy of selecting the *minimum*, m, of these, there should be no confusion. When multiple minima occurred for the smaller shift in outlier mean, $d = 2\sqrt{\chi^2_{0.975,2}}$, the drop in the value of (6) was less obvious. Again though, there should be no confusion if we simply choose the minimum of *any* minima occurring when $\alpha > 0$.

Figures 1-4 depict the sharp drop in our measure of Asymptotic Variance in the vicinity of the correct trimming proportion despite multiple minima.

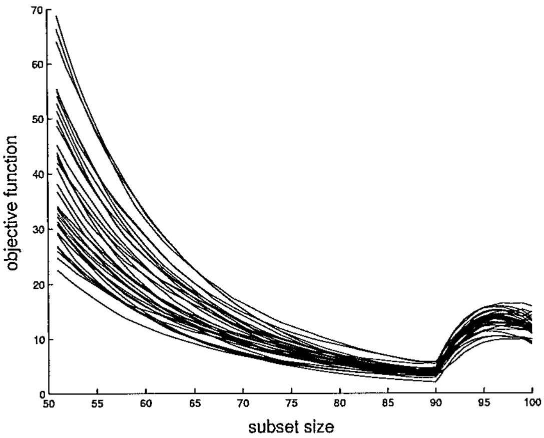

Figure 1: n=100, $\epsilon = 0.1$

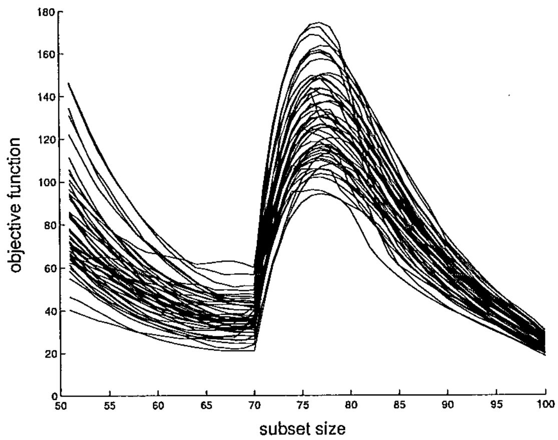

Figure 2: n=100, $\epsilon = 0.3$

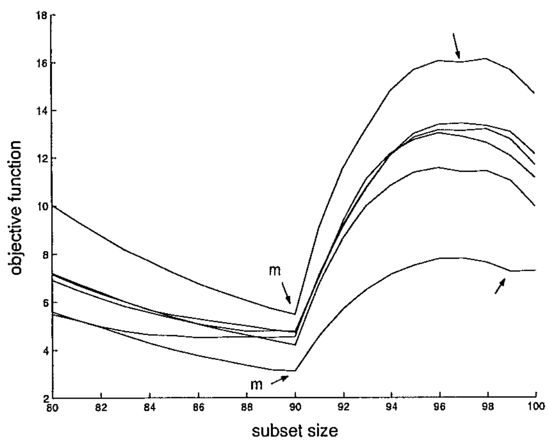

Figure 3: Excerpt from Figure 1

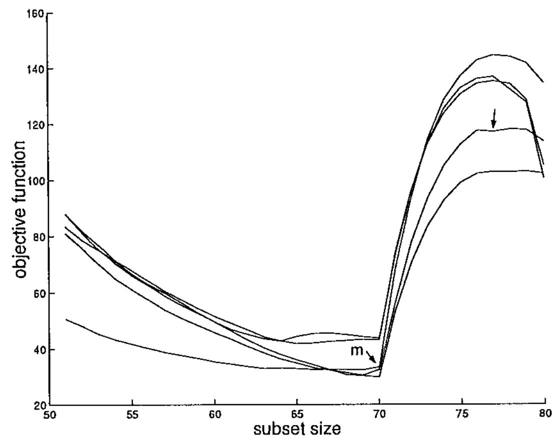

Figure 4: Excerpt from Figure 2

Table 3: Results for correlated data sets

$\rho = +0.5$	p_t	$\bar{\alpha}$	$\rho = +0.95$	p_t	$\bar{\alpha}$
n=20	0.070	0.0249	n=20	0.071	0.0225
n=50	0.0160	0.0029	n=50	0.010	0.0020
n=100	0.0030	< 0.0001	n=100	0.0030	< 0.0001

All simulations conducted thus far involved only those data sets, of dimension 2 and 4, generated randomly, it was decided to inspect our proposal Type 1, for samples of size $n = 20$, and Type 2 otherwise, on correlated bivariate data sets. The randomly generated *clean* data sets, for the results shown in Table 3, were thus converted into data sets of a pre-specified correlation, $\rho = 0.5$ and $\rho = 0.95$ respectively.

Table 3 contains similar results to Table 1, for clean data sets of dimension 2, for both pre-specified correlations. This is a crucial aspect of any outlier detection methodology, its measure of success needs to be independent of how any prospective data set may be correlated.

Applying the new proposal Type 1 to data sets of size $n \leq 30$ or Type 2 if the data set is of size $n > 30$ our algorithm is outlined more formally as follows:

Step 1: Calculate robust MCD estimate for location and scale, with $h = \lfloor \frac{n+p+1}{2} \rfloor$

Step 2: Order all observations ascending order of their Mahalanobis distance from this MCD estimate.

Step 3: For the subset of observations retained for the MCD estimate calculate the value of the objective function.

Step 4: Inflate this subset to include that observation within the complement of this subset possessing the smallest Mahalanobis distance from the centroid of this subset.

Step 5: Assess the value of the objective function for this inflated subset of observations.

Step 6: Re-order all observations in ascending order of their Mahalanobis distance with respect to this inflated subset, which may or may not result in the interchange of some members of this retained set and its complement.

Step 7: Repeat steps 4, 5 and 6 until all observations are included in the subset being inflated.

We take that α which corresponds to the minimum of *any* minima occurring for $\alpha > 0$ as the correct trimming proportion. If no local minima occur for $\alpha > 0$ then the data set is considered outlier free.

4 Comparison with other existing Algorithms

This algorithm has been compared with numerous methods, Rousseeuw & van Zomeren (1990), Hadi (1992, 1994), Rocke & Woodruff (1996) all involving a pre-specified outlier region, or fixed threshold. Cluster techniques for outlier identification have also been investigated, K-means (MacQueen 1967), Hierarchical (Mardia, Kent & Bibby 1979) and two more recent cluster methods based on a mixture likelihood, MINO (Rocke & Woodruff 1999) and the EM algorithm, EMCD (Coleman & Woodruffe 2000). Those methods identifying outliers as those observations beyond a specific threshold value had a tendency to consistently over trim and especially so for clean data sets.

It was Hadi's algorithm (1992, 1994), where a Forward Search algorithm in conjunction with a correction factor (Hadi 1994) was employed, that produced the more accurate results for outlier detection, in the presence of outliers. A drawback with Rocke & Woodruff (1996) was the need to derive cut off values using simulations for the specific sample and dimension size of any arbitrary data set. An adaptive algorithm identifies observations necessarily belonging to this outlier region without the need to specify what the cut off value is.

The cluster methods, K-mean and Hierarchical, required more extensive procedures, including the assessment of the existence of clusters and a measure of the partitioning imposed by any cluster configuration of the data (Rouseeeuw 1987). The more recently developed cluster techniques were most elaborate and have the ability of determining cluster "existence" and the number thereof but still had a tendency to over trim for *clean* data sets when used for outlier detection.

An alternative algorithm involving the adaptive approach, whence the threshold value is determined uniquely for each data set by the algorithm, is the Adaptive Reweighted Estimator (Gervini 2003). From various starting points, minimum volume ellipsoid (MVE), MCD and S-estimate for location and scale, this algorithm had a tendency to trim *clean* data sets even for sample sizes as high as $n = 500$. When our proposal, Type 2, was applied to clean data sets of size $n \geq 100$ trimming of data occurred very rarely, see Table 1. Hence our estimates for location and scale of clean samples were equal to the sample mean and covariance matrix. For sample sizes of $n \geq 50$ therefore, our relative MSE of location estimates and relative mean LCN's with respect to the *clean* sample were ≥ 0.98 and always 1.0 for samples $n \geq 200$.

In further Monte Carlo trials, when normally distributed samples, of dimension $p = 3,\ 10$, were corrupted with a proportion, $\epsilon = 0.1,\ 0.2$, of the data having its p-th variable distributed $N(k, 1)$ for $k = 1, \ldots, 20$. The maximum (with respect to k) median SE and LCN, as defined in Gervini (2003, p.131), with respect to an assumed location and LCN for an *uncontaminated* data set was recorded in Table 4.

Our maximum of the median standard errors regarding our location estimate occurred for $4 \leq k \leq 7$ and are slightly greater in comparison to Gervini (2003), for any $k \geq 10$ the outliers were nearly always identified, thus our results appeared to converge to the expected parameter estimates of a *clean* data set. Our median SE's for the simulations conducted for 10 dimensional data sets, with 20% of it shifted, were only slightly higher than that Adaptive Reweighted Estimate starting with the MCD estimate for location and scale. The results when using our proposal for the error in scale estimate, that is the values for the maximum median LCN, were again slightly greater than those for the Adaptive Reweighted Estimator whence applied to 3 dimensional samples, but were smaller for the 10 dimensional data sets.

With the proportion, $\epsilon = 0.2$, of data exhibiting a corrupted pth-variable, distributed

Table 4: Errors of location and scatter estimates for shifted normal

Error in location estimate				*Error in scale estimate*			
n	p	ϵ	maximum median SE	n	p	ϵ	maximum median LCN
50	3	0.1	0.1848	50	3	0.1	1.1136
		0.2	0.65			0.2	1.5225
	10	0.1	0.5134		10	0.1	2.408
		0.2	1.7089			0.2	2.9263
500	3	0.1	0.2191	500	3	0.1	1.1773
		0.2	0.9657			0.2	1.6418
	10	0.1	0.4666		10	0.1	1.8951
		0.2	1.9266			0.2	2.3977

Table 5: Errors of location and scatter estimates for amplified variance

Error in location estimate			*Error in scale estimate*		
n	p	median SE	n	p	median LCN
50	3	0.0604	50	3	0.6912
	10	0.2348		10	1.9005
500	3	0.0057	500	3	0.2174
	10	0.0239		10	0.5649

$N(0, 50)$, a contamination of identical location but amplified variance ensued and was also assessed under similar criteria, reporting the median standard error. The results, see Table 5, are excellent for $n = 500$. For $n = 50$ the results are as good as the results when the Adaptive Reweighted Estimate was applied to data sets with the same proportion, $\epsilon = 0.2$, of data shifted.

Our proposal, Type 2, was also applied to Cauchy distributed data of sizes $n = 50$ and $n = 500$ generated in both 3 and 10 dimensions for further comparison with Gervini (2003). Table 6 contains the relative median standard error of our estimate with respect to the Cauchy MLE for location and shows that out our proposal has yielded estimates for location very similar to the MLE.

Investigating the shape of our estimate for scale, we chose to calculate the relative median LCN with respect to the Cauchy MLE for the LCN echoing, again, the analysis of Gervini (2003). The new proposal estimate for scale produced relative values of $\approx 70\%$ to the median of the LCN's derived from the Cauchy MLE. These results for Cauchy data are as good as those found in Gervini (2003).

Table 6: Errors in Cauchy estimation with respect to Cauchy MLE

Error in location estimate			*Error in scale estimate*		
n	p	Relative median SE	n	p	Relative median LCN
50	3	0.9552	50	3	0.7072
	10	0.9298		10	0.6839
500	3	0.9354	500	3	0.7258
	10	0.9367		10	0.7069

Cauchy data sets are a good example of heavy tailed, symmetric distributions and if we proceed under the assumption of normality the data set will appear to consist of outliers. Substituting the Cauchy density for ϕ in (3), we can establish asymptotic minimums for (6) when applied to Cauchy data sets. Figures 5 and 6 show that our proposal Type 2 would trim, for both bivariate and trivariate Cauchy distributed samples, between $25 - 30\%$ of the data. This emphasizes that our proposal, Type 1 and Type 2, must be applied to data sets assumed normal.

Figures 7 and 8 depict an illustration of the asymptotic minimum of our objective function when applied to t-distributed data with 3 degrees of freedom. It is noticed that the trimming imposed by our proposal is not as severe, $3 - 5\%$.

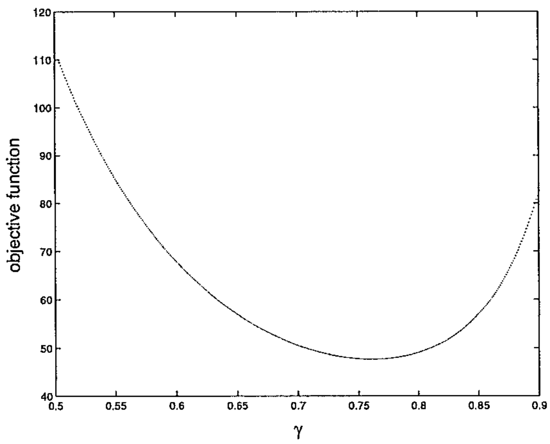

Figure 5: Bivariate Cauchy

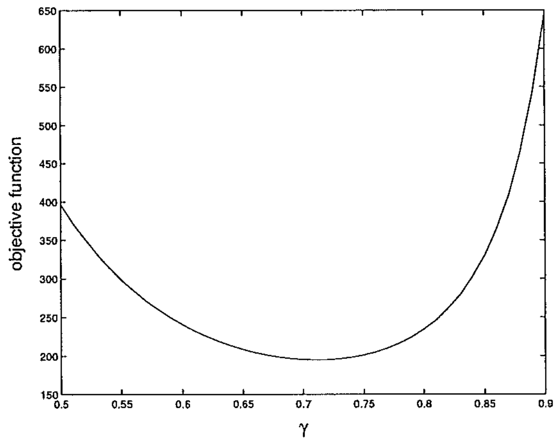

Figure 6: Trivariate Cauchy

Measure of asymptotic variances vs proportion γ of retained data distributed Cauchy

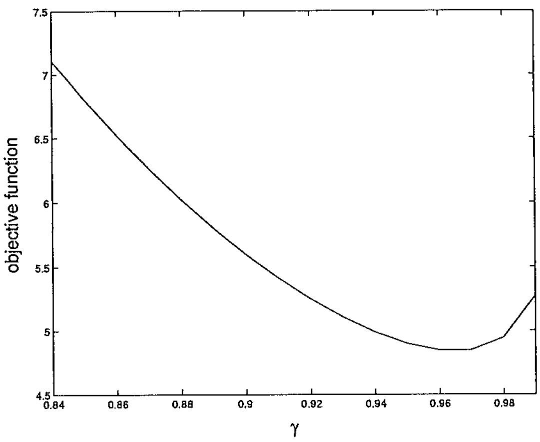

Figure 7: Bivariate t_3-distributed data

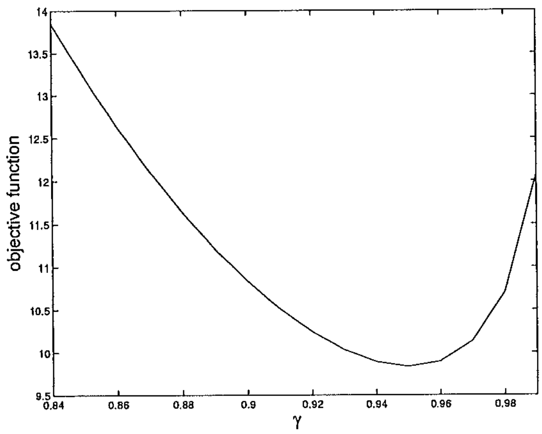

Figure 8: Trivariate t_3-distributed data

Measure of asymptotic variance vs proportion γ of retained t_3-distributed data

5 Real Data Sets

The first of our examples using real data sets concern the top 90 Career Batting Figures as of December 31st, 2003 for Australian and English test cricketers, see Table 7. The data set is 4 dimensional and has been chosen because of the extraordinary batting figures of Sir Donald Bradman. The new proposal, Type 2, seeks a subset of data corresponding to the minimum of any minima of (6) occurring for an $\alpha > 0$ unless the only minimum occurs at $\alpha = 0$. The 4 variables chosen in this analysis for outliers, were the number of innings played, x_1, number of hundreds, x_2, and fifties, x_3, scored and the number of runs amassed, x_4, by these top 90 batsmen.

When seeking subsets which minimize our objective function, (6), it transpires that there is a global minimum occurring at $\alpha = 1/90$ which corresponds to the subset of data with only Bradman's figures expelled. This satisfies our proposed method's criteria to consider

Table 7: top 90 batsmen

Name	x_1	x_2	x_3	x_4	Name	x_1	x_2	x_3	x_4
AR Border	265	27	63	11174	DL Amiss	88	11	11	3612
SR Waugh	258	32	49	10807	AW Greig	93	8	20	3599
GA Gooch	215	20	46	8900	AR Morris	79	12	12	3533
AJ Stewart	235	15	45	8463	EH Hendren	83	7	21	3525
DI Gower	204	18	39	8231	C Hill	89	7	19	3412
G Boycott	193	22	42	8114	GA Hick	114	6	18	3383
ME Waugh	209	20	47	8029	GM Wood	112	9	13	3374
MA Atherton	212	16	46	7728	FE Woolley	98	5	23	3283
MC Cowdrey	188	22	38	7624	KWR Fletcher	96	7	19	3272
MA Taylor	186	19	40	7525	ME Trescothick	81	5	21	3175
DC Boon	190	21	32	7422	VT Trumper	89	8	13	3163
WR Hammond	140	22	24	7249	AC Gilchrist	68	9	16	3159
GS Chappell	151	24	31	7110	MP Vaughan	71	10	8	3118
DG Bradman	**80**	**29**	**13**	**6996**	CC McDonald	83	5	17	3107
L Hutton	138	19	33	6971	AL Hassett	69	10	11	3073
KF Barrington	131	20	35	6806	KR Miller	87	7	13	2958
RN Harvey	137	21	24	6149	WW Armstrong	84	6	8	2863
DCS Compton	131	17	28	5807	GR Marsh	93	4	15	2854
RT Ponting	117	20	21	5749	KR Stackpole	80	7	14	2807
GP Thorpe	151	12	33	5552	NC O'Neill	69	6	15	2779
N Hussain	162	13	30	5430	M Leyland	65	9	10	2764
JB Hobbs	102	15	28	5410	GN Yallop	70	8	9	2756
KD Walters	125	15	33	5357	SJ McCabe	62	6	13	2748
IM Chappell	136	14	26	5345	C Washbrook	66	6	12	2569
MJ Slater	131	14	21	5312	GS Blewett	79	4	15	2552
WM Lawry	123	13	27	5234	BL D'Oliveira	70	5	15	2484
IT Botham	161	14	22	5200	DW Randall	79	7	12	2470
JH Edrich	127	12	24	5138	W Bardsley	66	6	14	2469
TW Graveney	123	11	20	4882	WJ Edrich	63	6	13	2440
JL Langer	116	16	20	4873	TG Evans	133	2	8	2439
RB Simpson	111	10	27	4869	LEG Ames	72	8	7	2434
IR Redpath	120	8	31	4737	MR Ramprakash	92	2	12	2350
AJ Lamb	139	14	18	4656	W Rhodes	98	2	11	2325
H Sutcliffe	84	16	23	4555	WM Woodfull	54	7	13	2300
PBH May	106	13	22	4537	DR Martyn	59	5	15	2292
ER Dexter	102	9	27	4502	TE Bailey	91	1	10	2290
KJ Hughes	124	9	22	4415	PJP Burge	68	4	12	2290
MW Gatting	138	10	21	4409	SE Gregory	100	4	8	2282
ML Hayden	83	17	13	4391	MJK Smith	78	3	11	2278
APE Knott	149	5	30	4389	SK Warne	146	0	8	2238
IA Healy	182	4	22	4356	R Benaud	97	3	9	2201
RA Smith	112	9	28	4236	CG Macartney	55	7	9	2131
MA Butcher	114	8	17	3790	WH Ponsford	48	7	6	2122
RW Marsh	150	3	16	3633	PE Richardson	56	5	9	2061
DM Jones	89	11	14	3631	RM Cowper	46	5	10	2061

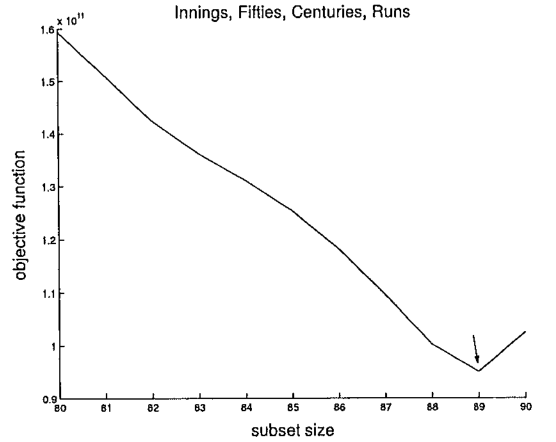

Figure 9: Minimum of (6) at $\alpha = 1/90$ when Bradman removed from the 4 dimensional data set in Table 7.

Bradman's figures as being a solitary outlier for this 4 dimensional data set. Figure 9 depicts this was the solitary minimum for $\alpha > 0$ over the subsets chosen by our Forward Search.

For a 3 dimensional example using these batting figures, we choose three related aspects when assessed together, that is innings played, fifties scored and runs amassed and can see the observation corresponding to Bradman marked with a 1 in Figure 10. The Bradman observation was again singled out by our Type 2 proposal as the solitary outlier even though from some angles, not shown, there was no way of pin-pointing his figures as outstanding in a 3-dimensional sense. Figure 11 depicts the drop in value of our objective function, (6), when Bradman's figures are removed from the 3 dimensional data set shown in Figure 10.

If we consider for a 2 dimensional example from this data set, fifties scored vs runs amassed, these two variables represent a fairly skewed data set where Bradman is marked with a 1 in Figure 12. The new proposal isolated this observation as the sole outlier which is most encouraging, observations 2 and 3 may also appear outlying to the naked eye but are consistent with the trend defined by the majority data.

Figure 13 is another good illustration of the impact on (6), the removal of Bradman's figures from Figure 12 had.

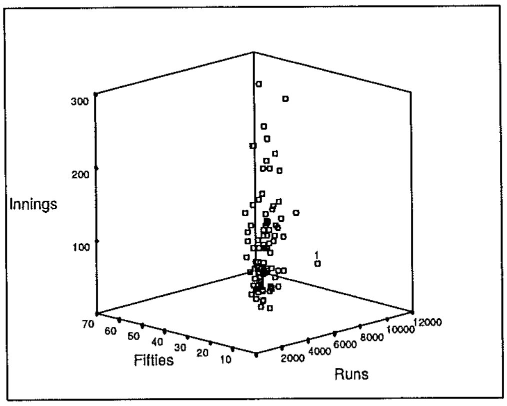

Figure 10: Innings vs Runs vs Fifties

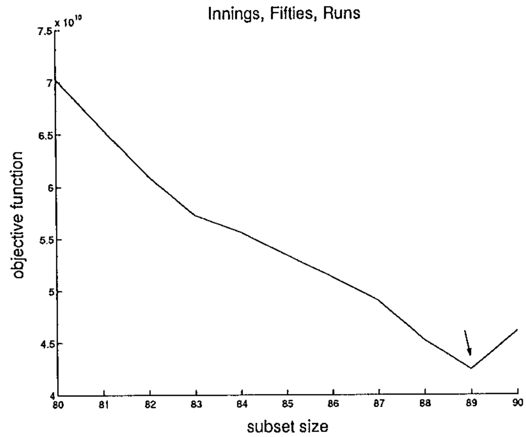

Figure 11: Minimum of (6) at $\alpha = 1/90$ when Bradman removed from the data set in Figure 10.

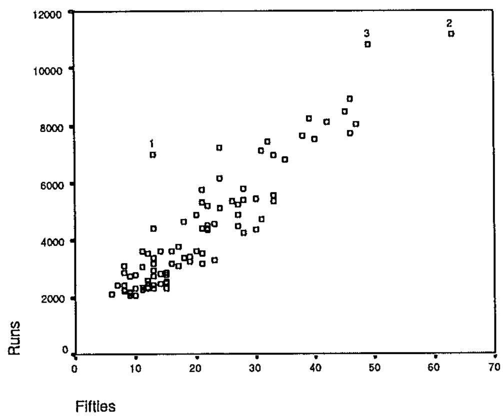

Figure 12: Runs vs Fifties

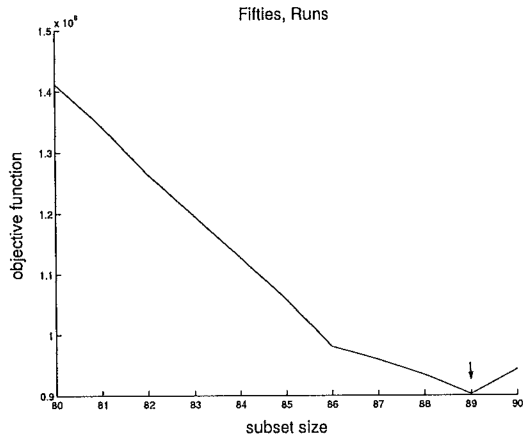

Figure 13: Minimum of (6) at $\alpha = 1/90$ when Bradman removed from the data set in Figure 12.

The following 3-dimensional plot, Figure 14, concerns a national sample of 6000 households with the main worker earning less than \$15,000 annually in 1966 (D.H. Greenberg and M. Kosters, Income Guarantees and the Working Poor, The Rand Corporation (R-579-OEO), December, 1970). These 6000 households were divided into 39 demographic subgroups for an analysis of the relationship between average asset holdings, average age, and average hourly wage. According to sources there exists one influential outlier at (wage=1.42, asset=1866, age=40.6) marked with the number 4. Our Type 2 proposal isolated four observations as outliers, those marked 1, 2, 3 and 4. Figure 15 displays the minimum of *two* minima occurring for $\alpha > 0$, hence the chosen trimming proportion corresponding to $\alpha = 4/39$.

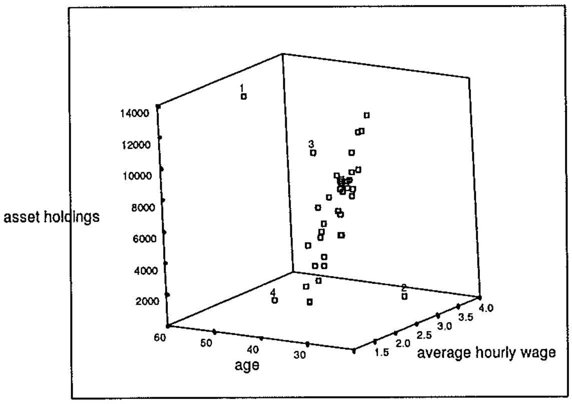

Figure 14: Wages and Hours

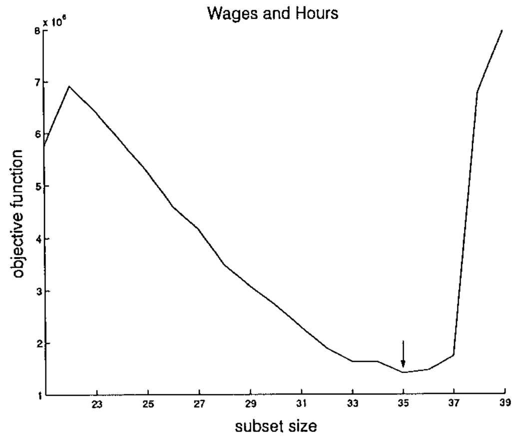

Figure 15: Smallest minimum for an $\alpha > 0$ at $\alpha = 4/39$

The last data set examined is contained in Table 8, it comprises of a 6-dimensional data set from Rousseeuw & Leroy 1987 (p 243), Wood Specific Gravity. According to an LMS regression fit and also ATLA in regression (cf. Clarke (2000)) observations 4,6,8 and 19 are outliers. The new proposal, Type 1, was applied treating the 5 independent variables, and the dependent variable, as a multivariate set of dimension 6. Only these 4 observations were identified as outlying and Figure 16 shows the sharp drop in the value of our objective function, (5), at $\alpha = 4/20$, the only minimum occurring for $\alpha > 0$.

Table 8: Modified Data on Wood Specific Gravity

Index	x_1	x_2	x_3	x_4	x_5	y
1	0.573	0.1059	0.465	0.538	0.841	0.534
2	0.651	0.1356	0.527	0.545	0.887	0.535
3	0.606	0.1273	0.494	0.521	0.92	0.57
4	**0.547**	**0.1135**	**0.531**	**0.519**	**0.915**	**0.548**
5	0.489	0.1231	0.562	0.455	0.824	0.481
6	**0.536**	**0.1182**	**0.592**	**0.464**	**0.854**	**0.475**
7	0.685	0.1564	0.631	0.564	0.914	0.486
8	**0.664**	**0.1588**	**0.506**	**0.481**	**0.867**	**0.554**
9	0.703	0.1335	0.519	0.484	0.812	0.519
10	0.653	0.1395	0.625	0.519	0.892	0.492
11	0.586	0.1114	0.505	0.565	0.889	0.517
12	0.534	0.1143	0.521	0.57	0.889	0.502
13	0.523	0.132	0.505	0.612	0.919	0.508
14	0.58	0.1249	0.546	0.608	0.954	0.52
15	0.448	0.1028	0.522	0.534	0.918	0.506
16	0.528	0.1057	0.424	0.566	0.909	0.568
17	0.437	0.1591	0.446	0.423	0.992	0.45
18	0.444	0.1628	0.429	0.411	0.984	0.431
19	**0.413**	**0.1673**	**0.418**	**0.43**	**0.978**	**0.423**
20	0.417	0.1687	0.405	0.415	0.981	0.401

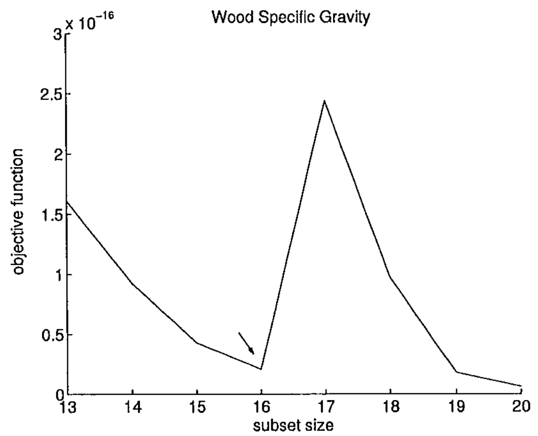

Figure 16: Solitary minimum for an $\alpha > 0$ at $\alpha = 4/20$

6 Conclusion

This paper has described a new proposal for the detection of outliers in multivariate data. It has been most successful in isolating both radial and clustered outliers with a single application. Along with its sensitivity to legitimate outliers, it has also been shown to be *insensitive* enough to rarely isolate an observation as outlying when it isn't. This technique simultaneously yields our parameter estimates for location and covariance, the normal theory maximum likelihood estimator calculated on the data set with the identified outliers removed. Code for our proposed algorithm was written in the R and Matlab computing packages and can be obtained by request from the authors.

7 Acknowledgements

We wish to thank Professor Ron Butler for sending us a copy of his notes pertaining to the paper Butler et al. (1993) and Professor Daniel Gervini for sending us a copy of his Matlab code for the multivariate Cauchy MLE. The authors are grateful to the referees for suggestions which improved the presentation of the paper. The first listed author is also grateful for prior collaboration with Professor Tadeusz Bednarski in relation to asymptotic theory for both univariate and regression estimation associated with trimmed likelihood estimation and its adaptive constructs.

References

ATKINSON, A. C. (1994). Fast very robust methods for the detection of multiple outliers. *J. Amer. Statist. Assoc.* **89**, 1329-1339.

ATKINSON, A. C.. RIANI, M. & CERIOLI, A. (2004). *Exploring Multivariate Data with the Forward Search*, New York, Springer.

BECKER, C. & GATHER, U. (1999). The masking breakdown point of multivariate outlier identification rules. *J. Amer. Statist. Assoc.* **94**, 947-955.

BEDNARSKI, T. & CLARKE, B. R. (1993). Trimmed likelihood estimation of location and scale of the normal distribution. *Austral. J. Statist.* **35**, 141-153.

BEDNARSKI, T. & CLARKE, B. R. (2002). Asymptotics for an adaptive trimmed likelihood location estimator. *Statistics* **36**, 1-8.

BUTLER, R. W. (1982). Nonparametric interval point prediction using data trimmed by a Grubbs type outlier rule. *Ann. Statist* **10**, 197-204.

BUTLER,R. W., DAVIES, P.L. & JHUN,M. (1993). Asymptotics for the minimum covariance determinant estimator. *Ann. Statist.* **21**, 1385-1400.

CLARKE, B. R. (1994). Empirical evidence for adaptive confidence intervals and identification of outliers using methods of trimming. *Austral. J. Statist.* **36**, 45-58.

CLARKE, B. R. (2000). An adaptive method of estimation and outlier detection in regression applicable for small to moderate sample sizes. *Discussiones Mathematicae: Prob. and Statist.* **20**, 25-50.

CLARKE, B. R., GAMBLE, D. K., & BEDNARSKI, T. (2000). A note on robustness of the β-trimmed mean. *Aust. N. Z. J. Stat.* **42** 113-117.

COLEMAN, D. A. & WOODRUFF, D. L. (2000). Cluster analysis for large datastes: An effective algorithm for maximizing the mixture likelihood. *J. Comput. Graph. Statist.* **9**, 672-688.

DAVIES, P. L. (1987). Asymptotic behaviour of S-estimators of multivariate location parameters and dispersion matrices. *Ann. Statist.* **15**, 1269-1292.

GERVINI, D. (2003). A robust and efficient adaptive reweighted estimator of multivariate location and scatter. *J. Mult. Anal.* **84**, 116-144.

HADI, A. S. (1992). Identifying multiple outliers in multivariate data. *J. Roy. Statist. Soc. Ser. B* **52**, 761-771.

HADI, A. S (1994). A modification of a method for the detection of outliers in multivariate samples. *J. Roy. Statist. Soc. Ser. B* **56**, 393-396.

JUAN, J. & PRIETO, F. J. (2001). Using angles to identify concentrated multivariate outliers. *Technometrics* **43**, 3; ABI/INFORM Global, 311-322.

MACQUEEN, J. (1967). Some methods for classification and analysis of multivariate observations, in: L. Le Cam & J. Neyman, Eds.. *Proc. 5th Berk. Symp. on Math. Statist. and Prob.* Berkeley, University of California Press, 281-297.

MARDIA, K. V., KENT, J. T. & BIBBY, J. M. (1979). *Multivariate Analysis.* London, Academic Press.

ROCKE, D. M. & WOODRUFF, D. L. (1993). Computation of robust estimates of multivariate location and shape. *Statist. Neerlandica.* **47** 27-42.

ROCKE, D. M. & WOODRUFF, D. L. (1996). Identification of outliers in multivariate data. *J. Amer. Statist. Assoc.* **91**, 1047-1061.

ROCKE, D. M. & WOODRUFF, D. L. (1999). A synthesis of outlier detection and cluster identification. unpublished manuscript. http://handel.cipic.ucdavis.edu/ dm-rocke/Synth5.pdf

ROUSSEEUW, P. J. (1983). Multivariate estimation with high breakdown point. In: Grossman, Pflug, Vincze and Wertz (Eds), *Mathematical Statistics and Applications*, Vol.B (1985) (Dordrecht:Reidel Publishing Co.), 283-297.

ROUSSEEUW, P. J. (1984) Least median of squares regression. *J. Amer. Statist. Assoc.* **79** 871-880.

ROUSSEEUW, P. J. (1987) Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *J. of Computat. & Appl. Math.* **20** 53-65.

ROUSSEEUW, P. J. & LEROY, A. M. (1987). *Robust Regression and Outlier Detection*, Wiley, New York.

ROUSSEEUW, P. J. & VAN ZOMEREN, B. C. (1990). Unmasking multivariate outliers and leverage points. *J. Amer. Statist. Assoc.* **85** 633-639.

WILKS, D. S. (1995) *Statistical Methods in the Atmospheric Sciences: An Introduction*, New York, Academic Press.

WOODRUFF, D. L. & ROCKE, D. M. (1993). Heuristic search algorithms for the minimum volume ellipsoid. *J. Computat. & Graph. Statist.* **2**, 69-95.